

Search Engines

Ashok Pillai



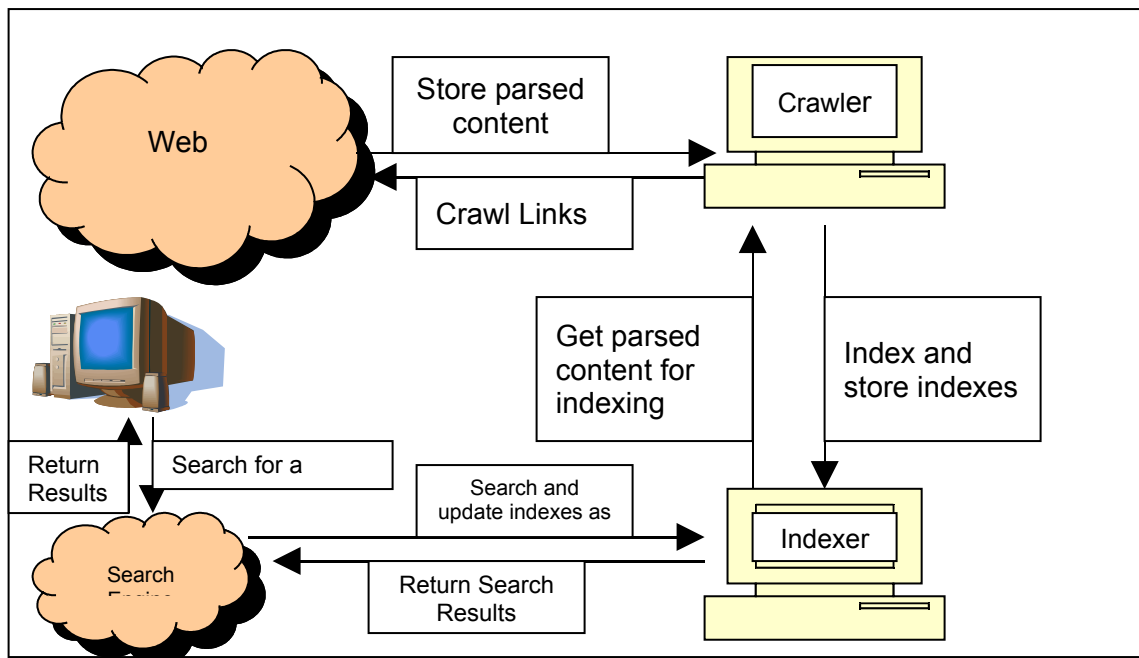
Search Engines

Search engines have become synonymous with a repository of knowledge. Though the data is from various sources across the Internet, the buzzword to find any information is 'Google'. Some other search sites have also done good but not as good as what 'Google' does and defines. There are even other search sites which target specific domains. "How things work", "Ask Jeeves", "Healthline" etc.

Let us understand the search engines more on how they work and the generic technology behind all of them.

A search engine has 3 basic components.

1. Web Crawlers
2. Indexers
3. Search Algorithm



Web Crawlers are programs or threads which crawl the internet to parse information out of web-sites. They are also called web spiders or web robots. They are scripts which conduct parsing and storing of information, by examining actual content of web-pages existing on the Web. They start off with a list of URLs, which are called seeds. As it proceeds through these URLs, it picks up HTML links from the parsed pages and appends to the list of seeds. Already accessed links are discarded. At the end of a crawling job, all the data collected is stored in a repository for indexing. We-crawling is done and frequent intervals to get updated data from web-sites. News based sites as well as blogging sites are crawled daily or at frequent intervals as they get updated very often. Other category of sites may be crawled once a month.

Indexers are complex programs which index all the data accumulated by the web crawlers. Search engines use text-based indexers which can index text so as each word is searchable. Before indexing certain stop-words are provided to the indexer. Stop-words are words which do not need to be searched viz. prepositions (if, the, and, then) etc Indexers do not index these stop-words. All text is listed against the URLs from where the text is found.

Indexing is done, whenever a web-crawling task is completed, on the new data accumulated. Indexing is a time-consuming and storage-centric task and the best indexers try to save on both time and space. Indexes can be stored in folders or in databases. Indexes can be arranged reflecting regions, domains, or any other broad categories to make search easier and faster.

Once the indexed data is available, comes the task of the **Search Algorithm**. When anyone uses a search engine to find a word, that word is searched against the indexes. These indexes are updated to reflect the number of times a link is opened by a click. Search algorithms differ on different search engines. This algorithm defines the efficacy of any search engine. Some of the parameters which affect the way results are shown after a search are following:

- a. **Ranking:** Whenever a search result is clicked or opened, the page opened count is incremented. This increases the ranking of the link for the text searched. Search results are ranked in this manner. There are many other ranking mechanisms prevalent.
- b. **Order of Search results:** Another aspect of search is the words provided for search. If it is a single word, e.g. 'run', will be searched as 'run', 'running', 'ran' etc. If 2 words are given then all the search which has both words in the same order comes first, then searches with both the words but not in any specific order and then for each word.
- c. **Boolean Searches:** Adding a + or AND in between two search words makes the algorithm find pages with both the words in it and no other. OR added between two words will search for any of the words
- d. **Wild-card searches:** Searching with wild card expressions or regular expressions are also supported on many search engines
- e. **Soundex Searches:** This helps in searching for words with sound similar to the word searched.